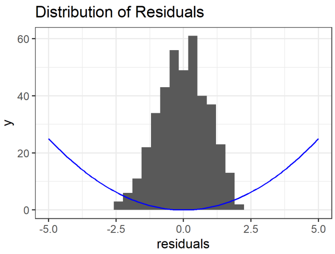
Introduction to Machine Learning

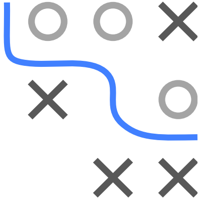
Advanced Risk Minimization
Maximum Likelihood vs. ERM



Learning goals

- Max. lik. and ERM are the same
- Gaussian errors = L2 loss
- Laplace errors = L1 loss
- Bernoulli targets vs. log loss

MAXIMUM LIKELIHOOD

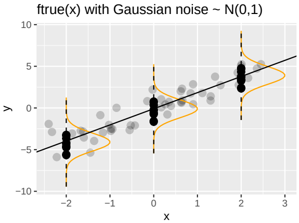


- Regression from a maximum likelihood perspective
- Assume data comes from $\mathbb{P}_{xy}$
- Conditional perspective:

$$y \mid \mathbf{x} \sim p(y \mid \mathbf{x}, \boldsymbol{\theta})$$

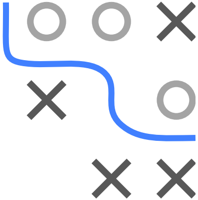
- Common case: true underlying relationship f_{true} with additive noise (surface plus noise model):

$$y = f_{\text{true}}(\mathbf{x}) + \epsilon$$



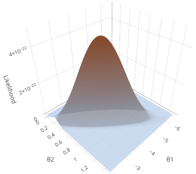
- f_{true} has params $\boldsymbol{\theta}$ and $\epsilon \sim \mathbb{P}_{\epsilon}$, with $\mathbb{E}[\epsilon] = 0$, $\epsilon \perp\!\!\!\perp \mathbf{x}$
- We now want to learn f_{true} (or its params)

MAXIMUM LIKELIHOOD



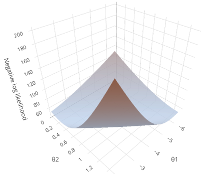
- Given i.i.d data $\mathcal{D} = \left(\left(\mathbf{x}^{(1)}, y^{(1)}\right), \ldots, \left(\mathbf{x}^{(n)}, y^{(n)}\right) \right)$ from $\mathbb{P}_{xy}$
- Max. likelihood maximizes **likelihood** of data under params

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} p\left(y^{(i)} \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}\right)$$

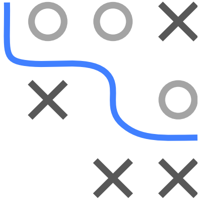


- Equivalent: minimize **negative log-likelihood (NLL)**

$$-\ell(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \log p\left(y^{(i)} \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}\right).$$



RISK MINIMIZATION



- In ML / ERM: instead of conditional distribution, pick a loss
- Our admissible functions come from hypothesis space
- But in stats, must assume some form of f_{true}, no difference
- Simply define neg. log-likelihood as **loss function**

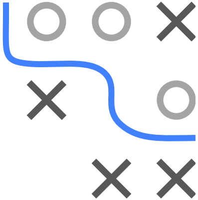
$$L\left(y, f(\mathbf{x} \mid \boldsymbol{\theta})\right) := -\log p(y \mid \mathbf{x}, \boldsymbol{\theta})$$

- Then, maximum-likelihood = ERM

$$-\ell(\boldsymbol{\theta}) = \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} L\left(y^{(i)}, f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)$$

- NB: When only interested in minimizer, we use $\propto$ as “proportional up to pos. multiplicative and general additive constants”

GAUSSIAN ERRORS - L2-LOSS



- Assume $y = f_{\text{true}}(\mathbf{x}) + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$
- Then $y \mid \mathbf{x} \sim \mathcal{N}(f_{\text{true}}(\mathbf{x}), \sigma^2)$ and likelihood is

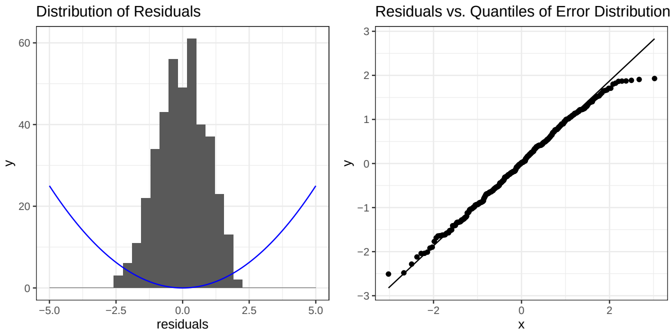
$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \prod_{i=1}^{n} p\left(y^{(i)} \,\middle|\, f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right), \sigma^2\right) \\
&\propto \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}\left(y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)^2\right)
\end{aligned}$$

- Minimizing Gaussian NLL is ERM with $L2$-loss

$$\begin{aligned}
-\ell(\boldsymbol{\theta}) &= -\log(\mathcal{L}(\boldsymbol{\theta})) \\
&\propto -\log\left(\prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}\left(y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)^2\right)\right) \\
&\propto \sum_{i=1}^{n} \left(y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)^2
\end{aligned}$$

GAUSSIAN ERRORS - L2-LOSS

- Simulate data $y \mid x \sim \mathcal{N}(f_{\text{true}}(x), 1)$ with $f_{\text{true}} = 0.2 \cdot x$
- Plot residuals as histogram, after fitting LM with $L2$-loss (blue)
- Compare emp. residuals vs. theor. quantiles via Q-Q-plot

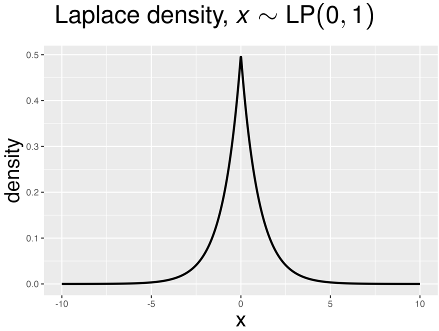


- Residuals are approximately Gaussian!

LAPLACE ERRORS - L1-LOSS

- Consider Laplacian errors ϵ, with density

$$\frac{1}{2\sigma}\exp\left(-\frac{|\epsilon|}{\sigma}\right), \sigma > 0$$



- Then

$$y = f_{\text{true}}(\mathbf{x}) + \epsilon$$

also follows Laplace distribution with mean $f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)$ and scale σ

LAPLACE ERRORS - L1-LOSS

- The likelihood is then

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \prod_{i=1}^{n} p\left(y^{(i)} \,\middle|\, f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right), \sigma\right) \\
&\propto \exp\left(-\frac{1}{\sigma} \sum_{i=1}^{n} \left| y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right) \right|\right)
\end{aligned}
$$

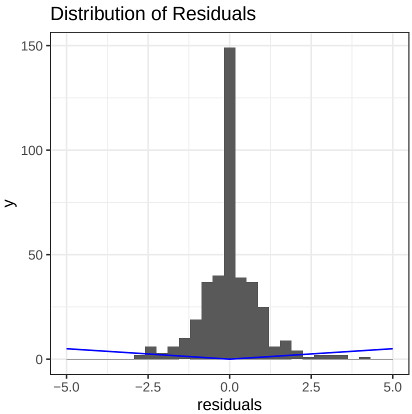
- The negative log-likelihood is

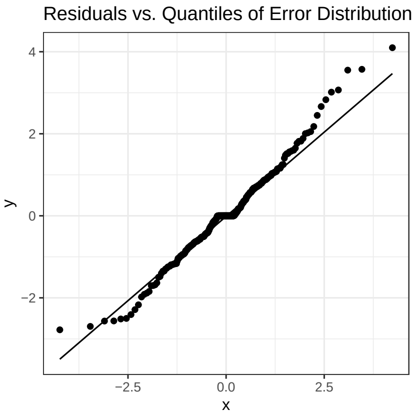
$$
-\ell(\boldsymbol{\theta}) \propto \sum_{i=1}^{n} \left| y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right) \right|
$$

- MLE for Laplacian errors = ERM with L1-loss
- Some losses correspond to more complex or less known error densities, like the Huber loss ▸ Meyer 2021
- Huber density is (unsurprisingly) a hybrid of Gaussian and Laplace

LAPLACE ERRORS - L1-LOSS

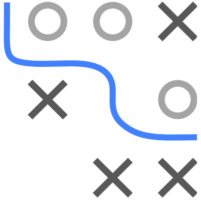
- Same setup, now with $y \mid x \sim \text{LP}(f_{\text{true}}(x), 1)$
- Now fit LM with L1 loss





- Again, residuals approximately match quantiles!

MAXIMUM LIKELIHOOD IN CLASSIFICATION



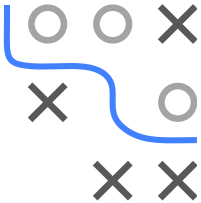
- Now binary classification
- $y \in \{0, 1\}$ is Bernoulli, $y \mid \mathbf{x} \sim \text{Bern}(\pi_{\text{true}}(\mathbf{x}))$
- NLL:

$$\begin{aligned} -\ell(\boldsymbol{\theta}) &= -\sum_{i=1}^{n} \log p\left(y^{(i)} \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}\right) \\ &= -\sum_{i=1}^{n} \log \left[\pi(\mathbf{x}^{(i)})^{y^{(i)}} \cdot (1 - \pi(\mathbf{x}^{(i)}))^{(1-y^{(i)})}\right] \\ &= \sum_{i=1}^{n} -y^{(i)} \log[\pi(\mathbf{x}^{(i)})] - (1 - y^{(i)}) \log[1 - \pi(\mathbf{x}^{(i)})] \end{aligned}$$

- Results in Bernoulli / log loss:

$$L(y, \pi(\mathbf{x})) = -y \log(\pi(\mathbf{x})) - (1 - y) \log(1 - \pi(\mathbf{x}))$$

DISTRIBUTIONS AND LOSSES



- For **every** error distribution $\mathbb{P}_\epsilon$, can derive an equivalent loss
- Leads to same point estimator for $\boldsymbol{\theta}$ as maximum-likelihood:

$$\hat{\theta} \in \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) \Leftrightarrow \hat{\theta} \in \arg\min_{\boldsymbol{\theta}} - \log(\mathcal{L}(\boldsymbol{\theta}))$$

- But **cannot** derive a pdf/error distrib. for every loss, e.g., Hinge loss; some prob. interpretation still possible ▸ Sollich 1999
- For dist.-based loss on residual $L(y, f(\mathbf{x})) = L_{\mathbb{P}}(r)$, ERM is fully equiv. to max. conditional log-likelihood $\log(p(r))$ if
  1. $\log(p(r))$ is affine trafo of $L_{\mathbb{P}}$ (undoing the $\propto$):
     $\log(p(r)) = a - bL_{\mathbb{P}}(r), \; a \in \mathbb{R}, b > 0$
  2. p is a pdf (non-negative and integrates to one)